Aspects Of The Predictive And Estimative Approaches In The Determination Of Probabilities

by

Seymour Geisser*

University of Minnesota

Technical Report No. 385

February 1981

Dedicated to the memory of Jerome Cornfield.

*This work was supported in part by NIH research grant GM-25271.

Aspects Of The Predictive And Estimative Approaches In The Determination Of Probabilities

Seymour Geisser*

University of Minnesota

Dedicated to the memory of Jerome Cornfield.

1. Introduction

In a previous paper, Geisser (1971), I argued that the predictive approach in statistics which had long been neglected or entirely disregarded, was a more appropriate vehicle for the transmission of a statistical inference, practically speaking, than the classical estimative approach. The predictive approach couches inferences and decisions in terms of observables or potential observables or interesting functions thereof while the estimative is involved with parameters. In this paper I shall attempt to present compelling arguments that predictivism or observablism is also a generally superior conceptual framework for inference which can subsume the estimative approach, when valid, as a limiting case, particularly in the determination of probabilities.

Predictive or observablistic inference is directed towards statements about a finite number of observables that conceptually have or had the potential of being generated. Clearly this includes sample surveys or any kind of sampling from a finite population since "estimation" here involves statements about some function of the finite totality of observations from the population which is often misdesignated as a "parameter." Passage to the limit will then include estimation -- since "parameters" will be entities that arise from the "prediction" of a function of an infinite number of potential observables.

Another situation where some difficulty in drawing a distinction between parameters and observables frequently arises, is in the measuring of some physical constant -- speed of light, length of a table, etc., with an imperfect

*This work was supported in part by NIH research grant GM-25271.

instrument i.e. subject to error. Here the model is that the observed value is the "parameter" or "true value" plus measurement error, say

$$X = \theta + e \ . \tag{1.1}$$

The true value θ is assumed to be a real physical entity observable up to some error of measurement and from a broad predictive point of view can be considered as an observable quantity. It can also be considered as a parameter whose value is associated with passage to the limit of a particular function of future observables $X_1, X_2 \ldots$. In these situations it is often appropriate to "estimate" θ rather than predict where a single or a finite number of new trials will occur. Here the value of prediction is restricted to the assessment of the measuring procedure. The determination of a distribution of values for θ will often be more useful than a predicting distribution for future values of X given previous trials. Passage to the limit of a predicting distribution of an appropriate function of future observations, however, would encompass the "estimation" of θ. But, aside from a few critical problems of this kind, most problems in statistics involve the evaluation or comparison of predicted or retrodicted values -- the latter addresses the inferral of past observations that could have occurred or occurred and perhaps were lost. At any rate, in these problems the variation in response to some agent is a real phenomenon in the sense that it is not measurement error and should not be represented by some true value plus error. However, the usual parametric estimative approach treats these problems as if there were a real θ that required estimation. A sensible analysis of these problems is to determine in some optimal or near optimal manner the distribution of response. This enables research workers to assess with reasonable accuracy how administration of such an agent will effect new patients. How the agent effects new patients

who, when certain known characteristics are taken into account, are more or less fungible with regard to the anticipated response to the agent, should be the goal of the analysis. If the finite future horizon consists of such patients fungible with those already observed and all to be treated similarly one would be interested in the probability that a proportion of them will be in some interval of response to an agent or in the entire distribution function itself. Under what circumstances would one be interested in the limiting case i.e. as the future patient number M grows? This would be sensible in two situations. First when there is difficulty in making the exact computation for a large M, the limiting case may serve as a convenient approximation. Secondly, conceiving of a hypothetically infinite number of potential patients may serve as a device for a normative evaluation of the response to an agent and, incidentally, provide a bound on the sharpest distinction among alternative therapies supplied by the data. In any event, as we shall later demonstrate, sensible parametric estimation problems can be formulated as special cases of observabilistic prediction by going to the limit i.e. as M goes to infinity.

In this paper I will take the Bayesian approach to demonstrate my point of view. I maintain that this approach allows making both the conceptual and the technical points in a much simpler manner than the classical Neyman-Pearson-Wald (NPW) frequentist approach. An analysis of the difference between these approaches in regard to this issue appears in Geisser (1980), where I showed that interpretation becomes problematic when the representation theorem of de Finetti (1937) is applied to exchangeable tolerance distributions. These distributions are the frequentist analogues of predictive distributions obtained in the Bayesian approach.

2. Parametric Estimation of Probabilities

In the usual Bayesian decision approach we have X the observation space, Θ the parameter space and A the action space with a random variable X having distribtuion $F_X(x|\theta)$, $\theta \in \Theta$, and a stipulated loss function $L(a(x),\gamma(\theta))$ where $a \in A$, γ is a mapping of Θ into Γ. Having observed $X = x$ and assumed a prior distribution $P(\theta)$ for θ, we compute the posterior density of θ,

$$P(\theta|x) = \int F_X(x|\theta) \quad dP(\theta) \tag{2.1}$$

where $P(\theta)$ is the prior distribution for θ.

The expected loss for each action **a** is

$$L(a) = \int L(a,\gamma)\ dP(\theta|x) = \int L(a,\gamma)\ dP(\gamma|x)\ . \tag{2.2}$$

where $P(\gamma|x)$ is the posterior distribution of γ and the conveyor for inferences about γ and upon which the loss function operates. The choice of the appropriate action, say a^*, is obtained as

$$L(a^*) = \min_a \ L(a). \tag{2.3}$$

We apply this now to the estimation of a probability function $\gamma(\theta) = Pr(X \in I|\theta)$ where I is some measureable set. In particular suppose $X^{(N)} = (X_1,\ldots, X_N)$ are i.i.d. random variables with common distribution function $F_X(x|\theta)$ where θ may be a set of parameters.

In principle then one can compute $P(\gamma|x^{(N)})$ from the distribution function of $P(\theta|x^{(N)})$. In appropriate circumstances the density of γ, $p(\gamma|x^{(N)})$ is most easily derivable from the posterior density of θ given $x^{(N)}$, i.e. from

$$p(\theta|x^{(N)}) \propto p(\theta) \prod_{i=1}^{N} f(x_i|\theta), \tag{2.4}$$

where $p(\theta)$ is the prior density of θ and $f(x|\theta)$ the sampling density. Hence one can compute $1 - \alpha$ probability limits on γ or point estimates of it.

This then would be the usual Bayesian approach to the estimation of a probability generated from the assumed underlying sampling distribution of the

observables. Although for many cases such a procedure could be quite complicated, we shall illustrate a case where it is fairly simple to achieve such an estimation program. Let $I = (y,\infty)$ so in this case

$$\gamma = 1 - F_X(y|\theta) = e^{-y\theta} \tag{2.5}$$

where F_X is the simple exponential distribution function. Further let $X_1,\ldots, X_d$ be fully observed values while X_j, $j = d+1,\ldots, N$ be censored at values x_j respectively. If we assume a non-informative prior $p(\theta) \propto \theta^{-1}$, and combine this with the likelihood

$$L(\theta) = \theta^d e^{-\theta N\bar{x}} \tag{2.6}$$

where $N\bar{x} = \sum_1^N x_i$, then the posterior density of θ is

$$p(\theta|\bar{x}) = (N\bar{x})^d \theta^{d-1} e^{-\theta N\bar{x}}/\Gamma(d) . \tag{2.7}$$

Hence $2\theta N\bar{x}$ is a χ^2 variate with $2d$ degrees of freedom. Now let $\gamma = e^{-y\theta}$ or $\theta = -y^{-1}\log\gamma$ so that the posterior density of γ is

$$p(\gamma|\bar{x}) = [\log \gamma^{-1}]^{d-1} \gamma^{y^{-1}N\bar{x}-1} \left(\frac{N\bar{x}}{y}\right)^{-d}/\Gamma(d) . \tag{2.8}$$

However for most applications we can actually use (2.7) instead of (2.8) because for $0 \leq a < b$,

$$Pr[a \leq \theta \leq b] = Pr[e^{-yb} \leq \gamma \leq e^{-ya}]. \tag{2.9}$$

So probability limits can be made in this case to depend on (2.7) and highest probability density intervals can, in principle, be computed as a single interval.

We also compute point estimators that minimize loss functions pointwise for each y. It is clear that the median $\tilde{\gamma}$ of γ can be obtained from (2.9) in view of the fact that the median $\tilde{\theta}$ of θ is given as

$$\frac{1}{2} = \Pr[\theta \leq \tilde{\theta}] = \Pr[\gamma \leq e^{-y\tilde{\theta}}] = \Pr[\gamma \leq \tilde{\gamma}] \tag{2.10}$$

thus $\tilde{\gamma} = e^{-y\tilde{\theta}}$ or since $\theta = \frac{\tilde{\chi}^2}{2N\bar{x}}$ where $\tilde{\chi}^2$ is the median of $\chi^2(2d)$ variate. Hence the median estimator of γ is

$$\tilde{\gamma} = \exp - \left(\frac{y\tilde{\chi}^2}{2N\bar{x}} \right) . \tag{2.11}$$

Here the loss function minimizes absolute error in terms of γ for each y. This Bayesian estimator will be close to the maximum likelihood estimator.

A second estimator $\bar{\gamma} = E(\gamma)$, which minimizes squared error in terms of γ for each y, may be computed as

$$\bar{\gamma} = \int \gamma(\theta,y)\ d\ P(\theta|x^{(N)}) = (1 + \frac{y}{N\bar{x}})^{-d} \tag{2.12}$$

which is obviously one minus the predictive distribution of a future observation drawn from this process. Thus this has a second interpretation.

The modal estimator γ_m, which minimizes in the limit a loss which is proportional to the length of the interval if correct and a constant less if incorrect, is obtained from maximizing the density $p(\gamma|\bar{x})$. This results in

$$\begin{aligned}
\gamma_m &= 1 && \text{for } y < N\bar{x},\ d = 1 \\
&= \tfrac{1}{2} \text{ (by definition)} && \text{for } y = N\bar{x},\ d = 1 \\
&= \exp[-\tfrac{(d-1)y}{N\bar{x}-y}] && \text{for } y \leq N\bar{x},\ d > 1 \\
&= 0 && \text{for } y > N\bar{x},\ d \geq 1 .
\end{aligned} \tag{2.13}$$

For further details on these estimators and comparisons with their frequentist counterparts see Geisser (1980).

3. The Predictive Approach

This approach essentially has a different goal, choosing an action which depends on the value of a future observable rather than on the value of the parameter. Here we postulate the sampling distribution of future $X_{(M)} = (X_{N+1}, \ldots, X_{N+M})$ the set of random variables conditional on θ and on a current set of random variables $X^{(N)} = (X_1, \ldots, X_N)$ to be

$$F(x_{(M)} | x^{(N)}, \theta) \ . \tag{3.1}$$

Instead of a parametric loss function we posit a predictive loss function c.f. Aitchison and Dunsmore (1975),

$$L_p(a, x_{(M)}), \tag{3.2}$$

which stipulates the loss incurred in taking action $a(x^{(N)}) = a$ upon having observed $X^{(N)} = x^{(N)}$ where $a \in A$ a space of actions, when $X^{(M)} = x_{(M)}$ occurs.

We obtain the predictive distribution of the future $X_{(M)}$ given the current data $X^{(N)} = x^{(N)}$ as

$$\overline{F(x_{(M)} | x^{(N)})} = \int F(x_{(M)} | x^{(N)}, \theta) dP(\theta | x^{(N)}) \tag{3.3}$$

from which all inferences and decisions about the future set of values, or any function of them, is obtained. The average predictive loss is

$$L_p(a) = \int L_p(a, x_{(M)}) \ dF(x_{(M)} | x^{(N)}) \ , \tag{3.4}$$

and the optimal a is that a^* such that

$$L_p(a^*) = \min_a L_p(a) \ . \tag{3.5}$$

In the case of i.i.d. random variables X_i, $i = 1, \ldots, N+M$ having common

distribution function $F_X(x|\theta)$, the predictive distribution of $X_{N+1},\ldots, X_{N+M}$ is

$$F(x_{N+1},\ldots, X_{N+M}|x^{(N)}) = \int [\prod_{i=1}^{M} F(x_{N+i}|\theta)] \quad dP(\theta|x^{(N)}) \tag{3.6}$$

represents a set of M exchangeable random variables (rarely independent) for each M. This results in an apparent distinction between the probability that a single future observation lies in some set I and the fraction of all such future observations that lie in I. Before we delve into that question we first indicate under what conditions it is appropriate to consider only the marginal distribution of a single future observation. Suppose that our loss function is of the additive form with equal loss for each component, a reasonable assumption when the future observations are essentially indistinguishable, then

$$L_p(a, x_{(M)}) = \sum_{i=1}^{M} L_p(a, x_{N+i}), \tag{3.7}$$

with average loss,

$$\begin{aligned} L_p^{(M)}(a) &= \sum_{i=1}^{M} \int..\int L_p(a,x_{N+i}) \; dF(x_{N+1},\ldots, x_{N+M}|x^{(N)}) \\ &= M\int L_p(a,x) \; dF(x|x^{(N)}) = ML_p(a) \end{aligned} \tag{3.8}$$

where $F(x|x^{(N)})$ represents the common marginal distribution of the exchangeable set of future random variables $X_{N+1},\ldots, X_{N+M}$. Hence the average loss depends only on the marginal distribution.

Given suitable conditions which permit interchanging an infinite number of integrals and an infinite sum then $M^{-1} L_p^{(M)}(a) = L_p(a)$ remains true for $M = \infty$.

As an example consider a loss function of the type

$$L(I_M, x_{(M)}) = \begin{cases} \delta \; V(I_M) & \text{if } x_{(M)} \notin I_M \\ \delta \; V(I_M) - K & \text{if } x_{(M)} \in I_M \end{cases} \tag{3.9}$$

for $\delta \geq 0$ and $K > 0$ where I_M is a countable number of non-overlapping M

dimensional intervals, and $V(I_M)$ is the volume of I_M. Then

$$L(I_M^*) = \min_{I_M} L(I_M) \tag{3.10}$$

yields as solution the set of non-overlapping M dimensional intervals I_M^* which results from the same set of non-overlapping intervals I^* for each coordinate of $x_{(M)}$, where

$$I^* = \{ x;\ f(x|x^{(N)}) > \frac{\delta}{K} \} \tag{3.11}$$

where $f(x|x^{(N)})$ is the common marginal density function. Hence it is essentially sensible to consider the same set I for each component of $X_{(M)}$ independently of the number of future observations under consideration.

In the light of this, consider the derived random variable

$$Y_i = \begin{cases} 1 & \text{if } X_{N+i} \in I \\ 0 & \text{otherwise} \end{cases} \tag{3.12}$$

for any particular measurable set I. Then $Y_1, \ldots, Y_M$ are exchangeable since $X_{N+1}, \ldots, X_{N+M}$ are, and

$$E(\bar{Y}) = \Pr(X_{N+i} \in I) = \int_I dF(x_{N+i}|x^{(N)}) = \Pr(Y_i=1) = q \tag{3.13}$$

where $\bar{Y} = M^{-1} \sum_{i=1}^{M} Y_i$. For $0 < q < 1$

$$\mathrm{Var}(\bar{Y}) = q(1-q)\left(\frac{1}{M} + \frac{M-1}{M} \rho \right) = q(1-q)\left(\frac{1-\rho}{M} + \rho \right), \tag{3.14}$$

where for all $i \neq j$, the common correlation coefficient is

$$\rho = \frac{\Pr[Y_i=1,\ Y_j=1] - q^2}{q(1-q)} . \tag{3.15}$$

Since ρ does not depend on M, $\rho \geq 0$, must clearly hold, otherwise $\mathrm{Var}(\bar{Y})$ will be negative. If the $Y_1, Y_2, \ldots$ are correlated (or even if only pairwise dependent) then

$$\lim_{M \to \infty} \mathrm{Var}(\bar{Y}) = \rho q(1-q) > 0 . \tag{3.16}$$

sufficient to guarantee the mutual independence of the X_i's, although it can be demonstrated for particular families of mixing distributions.

Because of exchangeability ρ in (3.14) is a function of N that tends to zero for increasing N independently of M under fairly general conditions, i.e., the conditions under which $P(\theta|x^{(N)})$ tends to concentrate all of its mass at a single point. But for any fixed N, the uncertainty in the fraction, $\bar{Y}$, of all future observations that lie in I, does not in general go to zero for finite N. It is true that its expectation is exactly the chance that $X_{N+i} \in I$ as given by the marginal predictive distribution of X_{N+i}.

In a previous paper, Geisser (1980) which also addressed this issue, I argued along the following slightly paraphrased lines. "Superficially one might think that $\bar{Y}$ should converge in probability to its mean value $q = Pr(X_{N+i} \in I)$ as M increases as it obviously would if the sequence $X_{N+1}, X_{N+2}, \ldots$ were independent instead of just exchangeable. And in point of fact in our original model the sequence is considered to be conditionally independent i.e. given θ. These points are not at all discordant in the sense that clearly, conditional on θ, $\bar{Y}$ converges to $\gamma(\theta) = Pr(X_{N+i} \in I|\theta)$ a constant and the fact that it doesn't unconditionally is a reflection of our uncertainty about the actual value of θ and the sampling distribution $F(x|\theta)$, e.g. when $I = (-\infty, x]$. It would be a paradox of the Bayesian procedure if unconditionally $\bar{Y}$ did converge to a constant since it would follow that $F(x_{N+i}|x^{(N)})$ must always be equivalent to the sampling distribution instead of it being the predictive distribution." Upon reflection I believe that this whole argument is perhaps better cast in terms of the difference between calculating the probability that the fraction of a finite number of potential observables that lie in a given set does not exceed a certain proportion and where the number is infinite.

In any event it is clear how de Finetti's theorem implies that the estimation of probabilities can be construed as a limiting case of prediction.

4. Illustration

We now illustrate the ideas of the previous section with the simple exponential distribution example of the section 2 where $I = (y,\infty)$ and $\gamma = e^{-\theta y}$. Then the probability that the fraction of M future X's greater than y is equal to r/M, is from (3.17),

$$\Pr[\bar{Y} = \frac{r}{M}] = \int \binom{M}{r} \gamma^r (1-\gamma)^{M-r} p(\gamma|x^{(N)})\, d\gamma \tag{4.1}$$

$$= \int \binom{M}{r} e^{-\theta y r} (1-e^{-\theta y})^{M-r} p(\theta|x^{(N)})\, d\theta,$$

where $p(\theta|x^{(N)})$ is given by (2.7). This yields

$$\Pr[\bar{Y} = \frac{r}{M}] = \binom{M}{r} (N\bar{x})^d \sum_{j=0}^{M-r} \binom{M-r}{j} (-1)^j [N\bar{x} + y(r+j)]^{-d} \tag{4.2}$$

and

$$\Pr[\bar{Y} \geq \frac{r}{M}] = \sum_{s=r}^{M} \Pr[\bar{Y} = \frac{s}{M}] . \tag{4.3}$$

Note that for $M = 1$

$$1 - \Pr[\bar{Y} \geq 1] = 1 - (1 + \frac{y}{N\bar{x}})^{-d} = F(y|x^{(N)}), \tag{4.4}$$

which is the predictive distribution of a single future observation. At the other extreme as M grows the distribution of $\bar{Y}$ tends to the distribution of γ.

There are several questions that are of interest at this point. If we are dealing with a small finite number of potential observations that are indistinguishable (in this case conditionally i.i.d.) we have no problem calculating predictive probabilities that involve them in some indistinguishable manner i.e. even more generally than r in I and $M-r$ not in I. This can be regarded as an appropriate consequence of assuming an additive loss function which is identical for each X_{N+i}. However, if we have a moderate or large number of potential observables to consider, it would be

of interest to determine the loss in accuracy in using the asymptotic result and if the accuracy is not sufficient for our purpose to derive approximations that are.

Certainly, if M is of some reasonable size the calculation of (4.2) or (4.3) is trivial.

If M is very large the double series computation becomes very burdensome and we might use the asymptotic result i.e. for $0 < a < b \leq 1$

$$\lim_{M\to\infty} \Pr[a \leq \bar{Y} \leq b] = \Pr[a \leq \gamma \leq b] \tag{4.5}$$

$$= \Pr[-y^{-1} \log b \leq \theta \leq -y^{-1} \log a]$$

$$= \Pr[-2d(\hat{\theta}y)^{-1} \log b \leq \chi^2_{2d} \leq -2d(\hat{\theta}y)^{-1} \log a] .$$

An approximation for intermediate values of M can be reasoned as follows: The set of random variables $X_{N+1}, \ldots, X_{N+M}$ are independent conditional on the single parameter θ hence unconditionally they are exchangeable as in (3.6). Then $M^{-1}(X_{N+1} + \cdots + X_{N+M}) = Z$, conditional on θ, has density

$$f_Z(z|\theta) \propto z^{M-1} e^{-\theta z}, \tag{4.6}$$

and the unconditional distribution of $(\hat{\theta}Z)^{-1}$, derived from the posterior distribtuion of θ given $\hat{\theta} = d/N\bar{x}$ of (2.7), is an F variate with $2d$ and $2M$ degress of freedom. Again by de Finetti's theorem, the random variable $W = Z^{-1}$ will tend to the random variable θ with density as gvien in (2.7). Hence as M increases $\bar{Y} \to \gamma = e^{-y\theta}$ and $W \to \theta$ then approximating the discrete distribution of $\bar{Y}$ by the continuous distribution of $V = e^{-yW}$ should be useful. This implies that

$$\Pr[a \leq \bar{Y} \leq b] \doteq \Pr[a \leq e^{-yW} \leq b] \tag{4.7}$$

$$= \Pr[-(\hat{\theta}y)^{-1} \log b \leq F_{2d,2M} \leq -(\hat{\theta}y)^{-1} \log a] .$$

Clearly a $(2d)^{-1}\chi^2_{2d}$ variate is on the whole more tightly concentrated about its central value than an $F_{2d,2M}$ variate so that for appropriately situated intervals (a, b) the probability for the inclusion of the F variate will increase as M grows with limiting probability determined by the former variate. This concords with the circumstance that the distribution of $\bar{Y}$ tends to be less concentrated than γ by virtue of the fact that they have a common mean

$$E(\bar{Y}) = E(\gamma) = \left(1 + \frac{y\hat{\theta}}{d}\right)^{-d} = 1 - F(y|x^{(N)}) \tag{4.8}$$

and

$$Var(\bar{Y}) = Var(\gamma) + M^{-1}\left[\left(1 + \frac{y\hat{\theta}}{d}\right)^{-d} - \left(1+\frac{2y\hat{\theta}}{d}\right)^{-d}\right] \tag{4.9}$$

where the coefficient of M^{-1} is non-negative with the $Var(\bar{Y})$ converging uniformly from above to the $Var(\gamma)$ as M grows.

As an example, consider the following data appearing in Gnedenko, Belyayev and Solovyev (1969, p. 176) consisting of a sample of $N = 100$ items tested, and time to failure recorded for each item until 500 standard time units (stu) have elapsed. The recorded failure times for 11 items were: 31, 49, 90, 135, 161, 249, 323, 353, 383, 436, 477. The other 89 items survived the test termination time so that the total time on test was $N\bar{x} = 47{,}187$. Suppose a lower bound on the future number of items out of M that will survive $y = 500$ stu with probability .95, is required. Solutions for r, given M, are to be obtained from the largest r such that

$$Pr[\bar{Y} \geq \frac{r}{M}] \geq .95 \tag{4.10}$$

The requisite computations from the data yield $\hat{\theta} = 11/47{,}187 = 2.33 \times 10^{-4}$ and $(\hat{\theta}y)^{-1} = 8.57945$. From (4.8) and (4.9) we obtain

$$E(\bar{Y}) = .891,$$
$$Var(\bar{Y}) = .0010 + 0065M^{-1}. \tag{4.11}$$

Table 1: Computation of Exact and Approximate values of $\Pr[Y \geq \frac{r}{M}]$ for Russian Data*

(M,r)	Exact $\Pr[\bar{Y} \geq \frac{r}{M}]$	F-Approximation $\Pr[F_{2d,2M} \leq -(\hat{\theta}y)^{-1} \log a]$	χ^2-Approximation $\Pr[\chi^2_{2d} \leq -2d(\hat{\theta}y)^{-1} \log a]$
(10,7)	.977	.977	1.000
(10,8)	.904	.771	.890
(20,15)	.972	.985	.999
(20,16)	.923	.917	.972
(30,23)	.973	.988	.998
(30,24)	.938	.953	.983
(30,25)	.870	.845	.897
(40,31)	.975	.989	.997
(40,32)	.949	.967	.987
(40,33)	.902	.909	.944
(50,40)	.956	.974	.989
(50,41)	.922	.938	.963
(50,42)	.867	.864	.897
(60,48)	.962	.979	.990
(60,49)	.935	.954	.972
(60,50)	.895	.906	.931
(70,56)	.966	.982	.990
(70,57)	.945	.963	.977
(70,58)	.913	.930	.949
(80,65)	.952	.969	.980
(80,66)	.927	.944	.960
(80,67)	.892	.904	.924
(90,73)	.958	.973	.983
(90,74)	.937	.954	.967
(90,75)	.909	.924	.940
(100,81)	.962	.977	.984
(100,82)	.945	.961	.972
(100,83)	.922	.938	.951

*$d = 11$, $a = (r + .5)/M$, $\hat{\theta} = 2.3312 \times 10^{-4}$, $y = 500$

In particular for $M = 100$, $S.D.(\bar{Y}) = .0326$, which would indicate that the search for the solution ought to be undertaken among the integers 82 to 84. The exact solution yields $r = 82$ with probability .945, for $r = 83$ the probability is .922.

The asymptotic value for (4.10) is

$$\Pr[\chi^2_{22} < -188.7479 \log a] = .95 \tag{4.12}$$

with solution obtained by equating the value to the right of the inequality in (4.12) to the $\chi^2_{22}(.95)$ percentage point. This yields $a = .835$, so that as an approximation to the finite $M = 100$ case above, making a correction for continuity, the lower bound is $r + \frac{1}{2} = 100a$ or $r = 83$. Application of the F approximation requires solving

$$\Pr[F_{22,200} < -8.57945 \log a] = .95 \tag{4.13}$$

and yields $a = .830$ as solution so that r is either 82 or 83. Hence the lower bound large sample approximation is about the same as the asymptotic value. Actually for $r = 82$ and $r = 83$ the F approximation yields the probabilities .961 and .938 respectively.

To have some idea of the usefulness of the approximation and the speed with which the exact probability tends to its asymptotic value, a table of these values is presented. In Table 1, the values of M are varied from 10 to 100 in increments of 10 and exact probabilities for r are computed which straddle .95. The F and χ^2 approximations to the exact values are also tabled.

The tabular entries give every indication that the F-approximation is, with few exceptions, closer to the exact probability than is the χ^2-approximation. Also, the F-approximation is always within a single unit (in r) of the integer closest to the .95 value. The exact probabilities are not difficult to program for the computer for moderate size M say up to 50. Beyond that special care must be taken in making the exact computations. In this latter range the F-approximation appears to yield values that are close enough to the exact values for most purposes. For each M and r the approximations should improve with increasing d.

Table 2 displays the exact values, the F and χ^2 approximations for $\Pr[\bar{Y} \geq .8]$ for varying M, keeping $\frac{r}{M} = .8$. It illustrates both the convergence to the asymptotic value (as M grows) and the deviation of the approximations from the exact value.

Table 2: Comparison of $\Pr[\bar{Y} \geq .8]$

M	Exact	F-Approximation	χ^2-Approximation
10	.904	.771	.890
20	.923	.917	.972
30	.938	.953	.983
40	.949	.967	.987
50	.956	.974	.989
60	.962	.979	.990
70	.966	.982	.990
80	.969	.984	.991
90	.972	.985	.991
100	.975	.986	.992
∞	.994	.994	.994

5. Acknowledgment

I am indebted to Murray Clayton for writing a program for the computation of the exact probabilities in Table 1. A copy of his program can be obtained upon request.

References

Aitchison, J. and Dunsmore, I.R. (1975). Statistical Prediction Analysis, Cambridge University Press, London.

Blum, J.R., Chernoff, H., Rosenblatt, M. and Teicher, H. (1958). Central limit theorems for interchangeable processes, Canadian J. Math., Vol 10, pp. 222-229.

de Finetti, B. (1937): La prévision: ses lois logiques, ses sources subjectives, Annales de l'Institut Henri Poincaré, 7, pp. 1-68.

Geisser, S. (1971). "The Inferential Use of Predictive Distributions," in Foundations of Statistical Inference, eds. V.P. Godambe and D.A. Sprott, Toronto: Holt, Rinehart and Winston, pp. 456-469.

Geisser, S. (1980). The estimation of distribution functions and the prediction of future observations, Proceedings of the Conference on Recent Developments in Statistical Methods and Applications, Academia Sinica, TAIWAN, pp. 193-208.

Gnedenko, B.B., Belyayev, Yu. K., and Solovyev, A.D. (1969). Mathematical Methods of Reliability Theory. Academic Press, New York and London.

Hewitt, E. and Savage, L.J. (1956). Symmetric measures on Cartesian products, Trans. Amer. Math. Soc., Vol. 80, pp. 470-501.